Identification de compatibilités entre tags descripteurs de lieux et apprentissage automatique

Estelle Delpech, Laurent Candillier, Léa Laporte, Samuel Phan

Identification de compatibilités entre descripteurs de lieux et apprentissage automatique

Estelle Delpech*,**, Laurent Candillier*,**, Léa Laporte*,***, Samuel PHAN*,**

*Nomao, 1 avenue Jean Rieux, 31 500 Toulouse, {prenom}@nomao.com, www.nomao.com
**Ebuzzing, 97 rue du Cherche-Midi, 75006 Paris, {prenom.nom}@ebuzzing.com, www.ebuzzing.com
***IRIT, Université de Toulouse 3, 118 route de Narbonne, 31062 Toulouse Cedex 9, {nom}@irit.fr, www.irit.fr

Résumé. Les travaux présentés dans cet article s'inscrivent dans le paradigme des recherches visant à acquérir des relations sémantiques à partir de folksonomies (ensemble de tags attribués à des ressources par des utilisateurs). Nous expérimentons plusieurs approches issues de l'état de l'art ainsi que l'apport de l'apprentissage automatique pour l'identification de relations entre tags. Nous obtenons dans le meilleur des cas un taux d'erreur de 23,7 % (relations non reconnues ou fausses), ce qui est encourageant au vu de la difficulté de la tâche (les annotateurs humains ont un taux de désaccord de 12%).

1 Introduction

Notre travail se situe dans le cadre d'un système d'agrégation de descriptifs de commerces. Dans ce système, chaque commerce est décrit (entre autres) par un jeu de tags. Des connaissances sur la compatibilité ou l'incompatibilité de ces tags peuvent aider à prendre la décision d'agréger ou pas deux descriptifs. Par exemple, un descriptif contenant le tag *fitness* ne peut pas être agrégé avec un descriptif contenant le tag *menuisier* car, *a priori*, un même commerce ne peut pas à la fois rendre des services de menuiserie et avoir un lien avec le fitness. Par contre, les tags *peinture* et *carrelage* sont compatibles car ils peuvent, par exemple, être associés à un magasin d'ameublement. Pour acquérir ces informations de compatibilité, nous disposons de données dont la structure est identique à celle d'une *folksonomie*. Les folksonomies sont le produit de systèmes d'annotation collaborative tels Flickr ou Delicious qui permettent à une communauté d'*utilisateurs* d'annoter manuellement des *ressources* (urls, photos) à l'aide de descripteurs (*tags*). Le but de notre étude est de tester diverses approches pour l'acquisition de relations entre tags à partir de folksonomies. L'approche retenue sera ensuite intégrée au système d'agrégation de descriptifs décrit plus haut. En plus de tester les approches état-de-l'art, nous explorons aussi l'apport de l'apprentissage automatique pour la détection de relations sémantiques à partir de folksonomies, ce qui, à notre connaissance, n'a pas encore été proposé.

L'article comprend 5 sections. La section 2 décrit les approches existantes pour l'identification de relations entre tags à partir de folksonomies. Les expériences menées sont décrites en section 3. Les données expérimentales et les résultats obtenus sont présentés en section 4. Les perspectives de travail sont discutées en section 5.

2 Travaux apparentés

Hotho et al. (2006b) définissent une folksonomie comme un 4-uplet $F := (U, T, R, Y)$ où $U = \{u_1, ...u_n\}$ est un ensemble d'utilisateurs, $T = \{t_1, ...t_m\}$ est un ensemble de tags, $R = \{r_1, ...r_p\}$ est un ensemble de ressources et $Y \subseteq U \times T \times R$. Un triplet $(u, t, r) \in Y$ correspond à l'attribution du tag t à la ressource r par l'utilisateur u. $T_{ur} := \{t \in T | (u, t, r) \in Y\}$ est l'ensemble des tags donnés à la ressource r par l'utilisateur u. Les travaux visant à identifier des relations entre tags à partir de folksonomies font état de trois types d'approches : statistiques, vectorielles et exploitation de ressources sémantiques [1].

Les approches statistiques sont les plus fréquemment employées. Elle portent sur la co-occurrence des tags, c'est-à-dire le nombre de fois où t_1 et t_2 ont été attribués à une même ressource par un même utilisateur, et qui est définie ainsi : $w(t_1, t_2) = |\{(u, r) \in U \times R | t_1, t_2 \in T_{ur}\}|$. Les mesures peuvent être asymétriques, comme dans le cas de Schmitz (2006) qui utilise la probabilité qu'une ressource ayant reçu le tag t_1 soit également annotée avec le tag t_2 ou symétriques comme la mesure Jaccard employée par Hassan-Montero et Herrero-Solana (2006) dans le but de construire une hiérarchie de concepts.

Les approches vectorielles consistent à représenter chaque tag dans un espace vectoriel dont les dimensions correspondent soit aux autres tags, soit aux ressources, soit aux utilisateurs. Par exemple, dans l'espace vectoriel correspondant aux ressources, chaque tag t est représenté par un vecteur $\vec{t} = \{w(t, r_1), ..w(t, r_{|R|})\}$ où $w(t, r) := |\{(t, u) \in T \times U | t \in T_{ur}\}|$. Les vecteurs sont comparés avec la mesure Cosinus. Ces approches ont été employées par Specia et Motta (2007) en association avec des connaissances issues d'ontologies en ligne dans le but de regrouper des tags en clusters et d'identifier des relations entre ces clusters. Heymann et Garcia-Molina (2006) les utilisent pour construire une taxonomie de tags.

Les approches exploitant des ressources sémantiques existantes consistent à projeter les tags de la folksonomie dans une ressource sémantique structurée en graphe [2] afin d'en déduire des relations sémantiques entre les tags - voir par exemple les travaux de Djuana et al. (2011). Cette technique a également été employée par Cattuto et al. (2008) ainsi que Markines et al. (2009) pour évaluer les approches vectorielles et statistiques. Pour cela, ils se basent sur la corrélation entre le score donné par la mesure à évaluer et la distance taxonomique de Jiang et Conrath (1997) calculée à partir de WordNet.

3 Expériences

Nos expériences ont consisté à tester les approches statistiques et vectorielles ainsi qu'à tester l'apport de l'apprentissage automatique. Nous avons utilisé deux types de ressources :

(i) un ensemble de triplets (u, t, r) correspondant à l'attribution d'un tag t au lieu r par la source Internet u. La structure de ces données est identique à celle d'une folksonomie, à ceci près que R est un ensemble de lieux (et non de documents) et que U est un ensemble de sources Internet décrivant les lieux (et non des utilisateurs). Toutes les approches testées s'appuient sur l'ensemble des tags (T) et l'ensemble des ressources (R) ainsi que

1. Plus anecdotiquement, d'autres travaux exploitent la structure de graphe des folksonomies et utilisent une version adaptée de l'algorithme PageRank (Hotho et al. 2006).

2. Par exemple : taxonomie, ontologie, thésaurus.

sur la fonction $\mathcal{R}(t) = \{r \in R | t \in T_{ur}, \forall u\}$ qui renvoie l'ensemble des lieux ayant reçu le tag t, quelles que soient les sources ayant attribué ce tag [3].

(ii) un arbre de tags initialement conçu pour la catégorisation de commerces (exemple de chemin : RACINE > *manger* > *restaurant* > *chic*). Les relations hiérarchiques ainsi que la distance entre deux tags sont potentiellement des indicateurs de leur compatibilité.

Cinq expériences ont été menées :

• **OVERLAP (approche statistique) :** OVERLAP est une mesure de type statistique dérivée de Jaccard. Elle correspond à un coefficient de chevauchement entre les lieux ayant reçu t_1 et les lieux ayant reçu t_2 :

$$Overlap(t_1, t_2) = \frac{|\mathcal{R}(t_1) \cap \mathcal{R}(t_2)|}{min(|\mathcal{R}(t_1)|, |\mathcal{R}(t_2)|)}$$

• **COSINUS (approche vectorielle) :** Dans cette approche, chaque tag t est représenté par un vecteur $\vec{t} = \{inter(t, t_1), ...inter(t, t_{|T|})\}$ où $inter(t, t_i) = |\mathcal{R}(t) \cap \mathcal{R}(t_i)|$. Ici, nous considérons que deux tags sont d'autant plus compatibles qu'ils ont des patrons de co-occurrences similaires (i.e. ils ont tendance à apparaître avec les mêmes tags). Par rapport à OVERLAP, cette approche a l'avantage de permettre de rapprocher deux tags même s'ils n'ont pas ou peu été attribués aux mêmes lieux. La similarité des vecteurs de deux tags est évaluée avec la mesure $Cosinus$:

$$Cosinus(\vec{t_1}, \vec{t_2}) = \frac{\sum_{i=1}^{|T|} inter(t_1, t_i) \cdot inter(t_2, t_i)}{\sqrt{\sum_{i=1}^{|T|} inter(t_1, t_i)^2} \cdot \sqrt{\sum_{i=1}^{|T|} inter(t_2, t_i)^2}}$$

• **ML_TAGTREE (apprentissage automatique à partir d'informations issue d'une ressource sémantique) :** Cette approche exploite des informations tirées de l'arbre de tags. À partir de cet arbre de tags, nous pouvons extraire, pour chaque paire de tags (t_1, t_2), 10 variables, décrivant soit des propriétés associées aux nœuds représentant les tags, soit des calculs de distance entre les tags :

1. nb. de chemins entre t_1 et t_2
2. distance min. entre t_1 et t_2 (nombre minimum d'arcs à parcourir pour relier t_1 à t_2).
3. distance max. entre t_1 et t_2 (nombre maximum d'arcs à parcourir pour relier t_1 à t_2).
4. nb. de chemins dans lesquels t_1 précède t_2 ou t_2 précède t_1
5. booléen indiquant si t_1 et t_2 sont à même distance de leur plus proche ancêtre commun
6. distance min. entre un des tags et leur plus proche ancêtre commun
7. distance max. entre un des tags et leur plus proche ancêtre commun
8. nb. de tags dans $\{t_1, t_2\}$ qui sont placés directement sous le nœud racine
9. nb. de tags dans $\{t_1, t_2\}$ correspondant à un nom de marque (i.e *Campanile, Ikéa...*)
10. nb. de tags dans $\{t_1, t_2\}$ ayant pour propriété de déclencher, lors de l'ajout du tag à un lieu, l'ajout automatique de tags plus génériques que lui [4]

3. La prise en compte du nombre de sources ayant attribué t à r est une information pertinente mais qui est, en l'état actuel du système, coûteuse à récupérer. Nous réservons son utilisation à des expériences ultérieures.

4. Cette information n'est pas calculée automatiquement, elle a été encodée lors de la création manuelle de l'arbre de tags.

Ces 10 variables ont été utilisées pour apprendre un modèle de classification à partir d'exemples de paires de tags annotées comme COMPATIBLE ou INCOMPATIBLE. L'apprentissage a été réalisé avec C5, un outil de génération d'arbre de décision [5] qui est une version améliorée de l'algorithme de Quinlan (1996).

• **ML_SIMPLE :** Dans cette expérience, nous utilisons toujours C5 auquel nous fournissons 4 variables très simples pour apprendre l'arbre de décision :$|\mathcal{R}(t_1) \cap \mathcal{R}(t_2)|$, $|\mathcal{R}(t_1) \cup \mathcal{R}(t_2)|$, $min(|\mathcal{R}(t_1)|, |\mathcal{R}(t_2)|)$ et $max(|\mathcal{R}(t_1)|, |\mathcal{R}(t_2)|)$.

• **ML_ALL :** Dans cette expérience, nous utilisons toujours C5, auquel nous fournissons toutes les informations utilisées dans les expériences précédentes. Le modèle de classification est donc appris à partir de 16 variables : la mesure $Overlap$, la mesure $Cosinus$, les 10 variables de ML_TAGTREE et les 4 variables de ML_SIMPLE.

4 Données expérimentales et résultats

Données Nous disposons de 15 millions de lieux, 3696 tags et 100 sources fournis par la société Nomao (`http://fr.nomao.com`). Pour apprendre les différents modèles de classification et évaluer nos expériences, nous avons constitué un jeu de 590 paires de tags annotées avec 2 classes : COMPATIBLE ou INCOMPATIBLE. Un tiers des paires a été annoté par au moins deux annotateurs [6]. Le taux de désaccord entre annotateurs est de 12%, soit un Kappa (Carletta, 1996) de 0,77. La répartition COMPATIBLE/INCOMPATIBLE est de 41%/59% respectivement.

Résultats Les approches ont été évaluées par validation croisée à 10 blocs. Pour les approches OVERLAP et COSINUS, un seuil de compatibilité a été appris en faisant varier sur les données d'apprentissage le seuil à partir duquel on considère que deux tags sont compatibles ; puis ce seuil a été évalué sur les données de tests. Nous avons comparé les résultats des approches deux à deux et pour chaque couple, nous avons appliqué le t-test unilatéral apparié. Nous considérons qu'une approche est significativement meilleure que l'autre si la valeur p du t-test est en-dessous de 5%. Nous obtenons les classements suivants (les valeurs p sont indiquées entre parenthèses) :

- ML_ALL est significativement meilleure que ML_TAGTREE ($p = 2\%$) et COSINUS ($p = 0.2\%$)
- ML_SIMPLE est significativement meilleure que COSINUS ($p = 1\%$)
- OVERLAP est significativement meilleure que COSINUS ($p = 2\%$)

Le tableau 1 indique la moyenne, médiane et l'écart-type des taux d'erreurs obtenus lors des 10 itérations. Nous observons que ML_SIMPLE obtient le taux d'erreur le plus stable.

Discussion Tout d'abord, nous notons que nous obtenons au mieux un taux d'erreur de 23,7%, que l'on peut considérer comme satisfaisant compte tenu de la difficulté de la tâche pour les humains (12% de désaccords). En effet, sur certaines paires de tags, il est très difficile de déterminer si les deux tags sont compatibles, par exemple : *hamburger* vs. *traiteur*, *concessionnaire* vs. *réparation de vélo*. Ensuite, vient le choix de l'approche qui sera finalement utilisée pour déterminer la compatibilité de deux tags. Même si ML_ALL est meilleure que COSINUS et ML_TAGTREE, elle est bien plus complexe à mettre en œuvre que deux autres approches au

5. `http://rulequest.com/download.html`

6. Nous avions 7 annotateurs en tout. En cas de désaccord, l'annotation retenue est celle de l'annotateur ayant le plus faible taux de désaccord avec les autres annotateurs.

	taux d'erreur moyen	taux d'erreur médian	écart-type
ML_ALL	0,237	0,229	0,052
ML_SIMPLE	**0,258**	**0,254**	**0,045**
OVERLAP	0,264	0,254	0,053
ML_TAGTREE	0,293	0,288	0,077
COSINUS	0,327	0,348	0,071

TAB. 1 – *Moyenne, médiane et écart-type des taux d'erreur obtenus par validation croisée*

taux d'erreur équivalent (ML_SIMPLE et OVERLAP) puisqu'elle nécessite le calcul de toutes les informations. Entre ML_SIMPLE et OVERLAP, nous retenons finalement ML_SIMPLE car c'est celle qui semble la plus robuste, au vu de la stabilité de son taux d'erreur.

5 Conclusion et perspectives

Nous avons présenté une étude visant à identifier des compatibilités entre descripteurs de lieux dans le but de l'intégrer à un système d'agrégation de descriptifs de lieux : la présence de tags incompatibles empêchera l'agrégation de deux descriptifs. Nous avons introduit l'utilisation de l'apprentissage automatique et nous avons testé trois modes de représentation des données : informations issues d'une ressource sémantique, un jeu de 4 indices "simples" et la combinaison des informations utilisées dans toutes nos expériences. Au final, c'est l'apprentissage automatique basé sur le jeu des 4 indices qui a été choisi car cette approche se place parmi les meilleures en termes de taux d'erreur et qu'elle est également la plus robuste.

La première perspective de travail est l'intégration de la compatibilité des tags dans le système d'agrégation et l'évaluation de son apport. Pour autant, notre méthode de détection de tags compatibles reste perfectible. Nous pensons à l'exploitation de ressources linguistiques plus riches que notre arbre de tags, comme par exemple le réseau lexical JEUX DE MOTS (Lafourcade, 2007) qui a l'avantage d'indiquer des relations lexicales entre mots (synonymie, hyponymie, etc.). Les autres perspectives de travail concernent la variation des différents paramètres de C5, l'essai d'autres approches d'apprentissage automatique (SVM, Naive Bayes...) ainsi que l'annotation de nouveaux exemples sélectionnés via des approches d'*apprentissage actif* (Settles, 2009).

Références

Carletta, J. (1996). Assessing agreement on classification tasks : The kappa statistic. *Computational Linguistics 22*(2), 249–254.

Cattuto, C., D. Benz, A. Hotho, et G. Stumme (2008). Semantic grounding of tag relatedness in social bookmarking systems. In *Proceedings of the 7th International Conference on The Semantic Web*, Karlsruhe, Germany, pp. 615–631.

Djuana, E., Y. Xu, et Y. Li (2011). Constructing tag ontology from folksonomy based on WordNet. In *Proceedings of the IADIS International Conference on Internet Technologies and Society 2011*, The East China Normal University, Shanghai, pp. In Press.

Hassan-Montero, Y. et V. Herrero-Solana (2006). Improving tag-clouds as visual information retrieval interfaces. In *Proceedings of the International Conference on Multidisciplinary Information Sciences and Technologies*.

Heymann, P. et H. Garcia-Molina (2006). Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical Report 2006-10, Standford InfoLab.

Hotho, A., R. Jäschke, C. Schmitz, et G. Stumme (2006a). FolkRank : a ranking algorithm for folksonomies. In *Proceedings of the joint workshop Lernen, Wissen und Adaptivität (Learning, Knowledge and Adaptability)*, Hildesheim, Germany, pp. 111–114.

Hotho, A., R. Jäschke, C. Schmitz, et G. Stumme (2006b). Information retrieval in folksonomies : search and ranking. In *Proceedings of the 3rd European conference on The Semantic Web : research and applications*, Budva, Montenegro, pp. 411–426.

Jiang, J. J. et D. W. Conrath (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the International Conference on Research in Computational Linguistics*, Taïwan.

Lafourcade, M. et A. Joubert (2010). Computing trees of named word usages from a crowdsourced lexical network. In *Proceedings of Computational Linguistics Applications*, Wisla, Polland, pp. 39–56.

Markines, B., C. Cattuto, F. Menczer, D. Benz, A. Hotho, et G. Stumme (2009). Evaluating similarity measures for emergent semantics of social tagging. In *Proceedings of the 18th international conference on World wide web*, Madrid, Spain, pp. 641–650.

Quinlan, R. (1996). Bagging, boosting and c4.5. In *13th National Conference on Artificial Intelligence*, pp. 725–730.

Schmitz, C. (2006). Inducing ontology from flickr tags. In *Proceedings of the Collaborative Web Tagging Workshop*.

Settles, B. (2009). Active learning literature survey. Computer Science Technical Report 1648, University of Wisconsin-Madison, Madison, USA.

Specia, L. et E. Motta (2007). Integrating folksonomies with the semantic web. In *Proceedings of the 4th European conference on The Semantic Web : Research and Applications*, Innsbruck, Austria, pp. 624–639.

Summary

The work presented in this paper deals with the extraction of semantic relations from folksonomies (set of tags attributed to resources by users). We experimented several state-of-the art methods as well as the added-value of machine learning for the identification of semantic relations between tags. We obtain an error rate of 23,7% (wrong or undetected relations), which is encouraging knowing that human annotators have a disagreement rate of 12%.